Who to Select: Identifying Critical Sources in Social Sensing

Dong Wang, Nathan Vance, Chao Huang

Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556

Abstract

Social sensing has emerged as a new data collection paradigm in networked sensing applications where humans are used as "sensors" to report their observations about the physical world. While many previous studies in social sensing focus on the problem of ascertaining the reliability of data sources and the correctness of their reported claims (often known as *truth discovery*), this paper investigates a new problem of *critical source selection*. The goal of this problem is to identify a subset of critical sources that can help effectively reduce the computational complexity of the original truth discovery problem and improve the accuracy of the analysis results. In this paper, we propose a new scheme, *Critical Source Selection (CSS)*, to find the critical set of sources by explicitly exploring both *dependency* and *speak rate* of sources. We evaluated the performance of our scheme and compared it to the state-of-the-art baselines using two data traces collected from a real world social sensing application. The results showed that our scheme significantly outperforms the baselines by finding more truthful information at a higher speed.

Keywords: Source Selection, Source Dependency, Speak Rate, Social Sensing, Twitter

1. Introduction

This paper develops a new scheme to solve the critical source selection problem in social sensing applications. Social sensing has emerged as a new net-

worked sensing paradigm of collecting observations about the physical environment from humans or devices on their behalf. This paradigm is motivated by the proliferation of digital sensors in the possession of individuals (e.g., smartphones) and the wide adaptation of online social media (e.g., Twitter, Facebook). In social sensing applications, people can report certain observations about their environment such as traffic conditions at various locales [1], pothole information on streets [2], and available gas stations in the aftermath of a disaster [3]. One key challenge of using "humans as sensors" is to estimate the correctness of observations (i.e., *claims*) and the reliability of data sources without knowing ground truth about the situation *a priori*. We refer to this problem as the *truth discovery problem*.

In this paper, we study a new problem of *critical source selection* where the goal is to identify a subset of critical sources that can reduce the computational complexity of the original truth discovery problem and improve the accuracy of the analysis results. First, it is critical to consider the source dependency in solving this problem. In social sensing, it is not unusual for a human source to forward claims they received from others (e.g., friends from their social networks) [4]. Figure 1 shows some simple examples extracted from real-world Twitter data where sources with social connections (i.e., following relationship) report the same claim. From a networked sensing perspective, such dependencies between sources can easily introduce correlation and redundancy between reported observations, which are shown to affect truth discovery results negatively if they are not appropriately modeled [5]. Previous works [6, 7, 5, 8] have started to account for dependencies between sources in truth discovery tasks by partitioning them into independent groups where sources in different groups are considered to be independent. However, the complexity of their solutions grow exponentially with respect to the maximum size of the independent groups, making them impractical in many large-scale social sensing applications [6]. In this paper, we develop a new source selection scheme to explicitly consider the source dependency in the source selection process.

In addition to the source dependency, the speak rate of a source (i.e., how

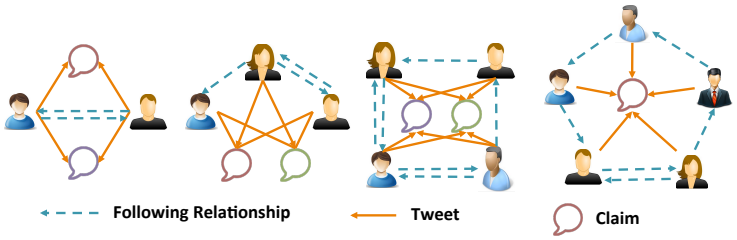


Figure 1: Source Dependency Examples on Twitter

chatty a source is) is another important factor to consider in the critical source selection solution. In social sensing, different sources often report different numbers of claims. The speak rate of a source has a strong positive correlation with both the accuracy and the granularity of the source reliability estimation, which also directly affects the estimation of the claim correctness [9]. Therefore, the goal of our critical sensor selection scheme is to (i) maximize the average speak rate of the selected sources and (ii) minimize the dependency between them. However, those two objectives can be at odds with each other, which makes the critical sensor selection problem non-trivial to solve.

Previous work has made significant progress towards source selection in sensor network and data fusion [10, 11, 12, 13, 14, 15]. However, most current solutions ignore either the source dependency or the speak rate in their models, which has led to suboptimal source selection results because redundant sources or sources with inaccurate source reliability estimations are selected. In this paper, we present a Critical Source Selection (CSS) scheme that explicitly incorporates both the *source dependency* and the *speak rate* feature into the critical source selection process. In particular, we formulate our critical source selection problem as a constraint optimization problem with multiple objectives, and we develop an efficient algorithm to solve it. We evaluate our CSS scheme in comparison with the state-of-the-art baselines using two real-world social sensing data traces collected from Twitter (i.e., the Paris Attack event in 2015 and the Oregon Shooting event in 2015). The results show that our scheme significantly outperforms the baselines by finding more truthful information at

a faster speed.

In summary, our contributions are as follows:

- We investigate the problem of critical source selection in social sensing to reduce the complexity of the truth discovery problem while simultaneously improving the accuracy of estimation results.
- We develop a new approach (CSS) that selects a critical set of sources by exploring both their source dependencies and their speak rates.
- We perform extensive experiments to compare the performance of our CSS scheme with state-of-the-art baselines using real-world social sensing data. The evaluation results demonstrate the effectiveness and efficiency of our scheme.

A preliminary version of this work has been published in [16]. This work significantly expands on our previous work and makes new contributions as follows. *First*, we extend our previous proposed model by developing a new annealing based process to increase the probability of reaching a globally optimal solution (Section 4). *Second*, we formally prove that the critical source selection problem in our work is NP-hard (Section 4). *Third*, we compare our scheme with more recent baselines using the real-world datasets and carry out a more comprehensive evaluation and comparison between the CSS scheme and the state-of-the-art techniques (Section 5). *Fourth*, we perform a set of new experiments to investigate the effect of parameters in the proposed model and study the robustness of the model with respect to the changes of the parameters (Section 5). *Finally*, we also evaluate the execution time of all compared algorithms to study their computational efficiency (Section 5).

The rest of this paper is organized as follows: we discuss the related work in Section 2. In Section 3, we present the problem of critical source selection. The proposed critical source selection scheme is discussed in Section 4. Experiment and evaluation are presented in Section 5. Finally, we conclude the paper in Section 7.

2. Related Work

Social Sensing. has emerged as a new sensing paradigm which attracted much attention in sensor networks research [17], urban sensing [18], surrogate sensing [19], Internet of Things [20], and data distillation [21]. The key idea of social sensing is to use humans as sensors in many sensing applications such as participatory sensing [22] and opportunistic sensing [23]. In particular, human sensors can contribute their observations through "sensing campaigns" [24] or social data scavenging [25]. Current works in social sensing have addressed important challenges in many relevant fields such as privacy perseverance [26], truth estimation [27], social signal processing [28], social sensor profiling [29], semantics of the sensing content [30, 31], and social interaction promotions [32]. However, source selection remains a critical and open research question in social sensing. In this work, we study the problem of *critical source selection* to reduce the computational complexity of the truth discovery problem and improve the accuracy of the analysis results.

Truth Discovery in Social Sensing. Data quality and trustworthiness is a fundamental challenge in social sensing. Prior works in social sensing have made significant advances to infer the credibility of reported data [33, 6, 34, 7]. For example, Ouyang et al. [33] investigated the potential of leveraging crowds as sensors to detect the true value of quantitative characteristics from noisy social sensing data. Huang et al. explored the *topic relevance of claims and arbitrary source dependency problem* in social sensing and developed a topic-aware truth discovery solution [6]. Zhang et al. developed a reliable truth discovery solution that is robust to sparse data and misinformation in social sensing [35]. Zhao et al. studied the problem of real-time truth discovery and developed a probabilistic model to efficiently handle streaming data [34]. Wang at al. considered source dependency by assuming that it can be represented by sets of disjoint trees [7]. All the above works solve the *truth discovery problem* and focus on modeling the relationship between source reliability and claim correctness. In contrast, this paper solves a new problem of *critical source*

selection which can help improve both the effectiveness and the efficiency of the above truth discovery solutions.

In addition, a few recent truth discovery solutions focus on improving efficiency by using streaming approaches [36, 37, 38]. For example, Wang et al. developed a streaming truth discovery scheme to recursively update the estimation results by leaveraging the previous estimation and the CRLBs of the estimation [36]. Zhang et al. proposed another category of streaming truth discovery approaches by explicitly addressing the scalability and physical constraints in social sensing application [37, 38]. However, the above works did not consider the critical source selection problem in their truth discovery solutions. In sharp contrast to those works, this paper improves the efficiency of the truth discovery solutions by solving a new critical source selection problem.

Source Selection in Social Sensing. There exists a good amount of work on the topic of *source selection* in networked sensing, data mining, and machine learning communities [10, 11, 12, 39, 13]. For example, Uddin et al. investigated the problem of diversifying the source selection in social sensing based on the social connections between sources. Rekatsinas et al. [11] studied the problem of source selection for dynamic sources whose contents change over time. Dong et al. [12] proposed an algorithm to select a subset of sources in data fusion applications by considering integration cost. Hosseini et al. selected the subset of data sources to predict the state of all other sources by considering source correlations [13]. Amintoosi et al. [39] proposed a a privacy-aware participant selection framework that explicitly protects users' privacy in the social sensing applications. However, most current solutions ignore either the source dependency or the speak rate in their models. In contrast, this paper explicitly incorporates both the *source dependency* and *speak rate* into the critical source selection process.

Hybrid Sensing. Finally, our work is also related to the hybrid sensing work [40, 41, 42]. For example, Zhang et al. developed a hybrid sensing system to study the localization problem for mobile robot in indoor environments [40]. Niforatos et al. proposed a crowdsourcing platform to study the data fusion

problem on weather related sensory data [41]. Subari et al. leveraged Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) methods to conduct comparisons between fusion methods [42]. The results of this work can be complementary to the above applications because selected critical sources normally lead to a better understanding of data fusion strategies in hybrid sensing by exploring both the source dependency and their speak rates.

3. Problem Formulation

We consider a social sensing scenario where a set of X sources (denoted as S) jointly report a set of Y claims (denoted as C). We denote an individual source as $S_i \in S,\ i \in [1, ..., X]$ and an individual claim as $C_j \in C,\ j \in [1, ..., Y]$, where i and j are the source and claim index respectively. The same claim can be made by multiple sources and each source can report multiple claims. We define the following terms we will use in our problem formulation.

Definition 1. ***Source-Claim Matrix*** *SC. We define the Source-Claim Matrix* $\boldsymbol{SC}_{X \times Y}$ *to represent whether a source reports a claim or not. In particular, in SC, we set* $SC_{i,j} = 1$ *if source* S_i *reports claim* C_j *and* $SC_{i,j} = 0$ *otherwise.*

Definition 2. ***Speak-Rate Vector*** *SR. We define the Speak-Rate Vector* $\boldsymbol{SR}_i$ *to represent how chatty a source is. Specifically, the element* SR_i *in SR is the number of claims reported by source* S_i *normalized by the total number of claims:* $SR_i = \frac{\Sigma_{j=1}^{Y} SC_{i,j}}{Y}$.

Definition 3. ***Source-Dependency-Score Matrix*** *SDS. We define the Source-Dependency-Score Matrix* $\boldsymbol{SDS}_{X \times X}$ *to represent dependency between each pair of sources. Specifically, the element* $SDS_{i,i'}$ *in SDS is the number of common claims reported by both source* S_i *and* $S_{i'}$.

We summarize the defined notations in Table 1.

In social sensing applications, the estimation accuracy of a source's reliability is positively correlated with the speak rate of the source [43]. The first objective of our critical source selection problem is to maximize the speak rates of the

Table 1: Summary of Notations

Symbol	Interpretation
S	set of sources
C	set of claims
SC	source-claim matrix
SR	speak-rate vector
SDS	source-dependency-score matrix

set of selected sources. Furthermore, observations from independent sources often provide more critical information to solve the truth discovery problem [7]. In Definition 3, we use the number of commonly reported claims to measure the dependency between two sources. This is based on the assumption that two independent sources are less likely to report many claims in common [6]. Therefore, the second objective here is to minimize the dependencies among the selected sources. Finally, the claims reported by the selected sources should cover all claims in C for the completeness of the problem.

Also note that we assume that the claims reported by the selected sources should cover all claims in C. The reason is we do not want to leave the correctness of any claims undecided because the source(s) who report that claim are removed. If a claim is not reported by many sources, then it is likely to be classified as a false claim by the truth discovery solutions. It does not hurt to include that claim in our problem setting. Also, based on the results from previous literature [44], some true claims can be reported by a small number of reliable sources. If we can correctly infer the correctness of a subset of claims without running the truth discovery algorithms (e.g., by analyzing the corroboration from different sources on the claim), we can actually remove this subset of claims from C to improve the efficiency of our CCS scheme.

With the above definitions, we can formulate the *critical source selection* problem as follows: given the Source-Claim Matrix $\boldsymbol{SC}_{X\times Y}$, Speak-Rate Vector $\boldsymbol{SR}_X$ and Source-Dependency-Score Matrix $\boldsymbol{SDS}_{X\times X}$, the goal is to select

the set of critical sources (denoted by S^*) whose reported claims cover the claim set C while maximizing their total speak rates and minimizing their total dependency scores. Formally, the problem can be represented as follows:

$$\begin{aligned} & max \sum_{i=1}^{X} SR_i \cdot \delta_i \\ & min \sum_{i=1}^{X} \sum_{i' \neq i} SDS_{i,i'} \cdot \delta_i \cdot \delta_{i'} \\ & s.t. \;\; \delta_i \in \{0,1\}, \quad i = 1, ..., X \\ & \quad \bigcup C_{S_i} = C, \quad S_i \in S \text{ where } \delta_i = 1 \end{aligned} \tag{1}$$

where C_{S_i} represents the set of claims asserted by source S_i and $\delta_i = 1$(or 0) indicates that source S_i is selected (or not).

In the above problem formulation, we cast our critical source selection problem as a multi-objective optimization (or Pareto optimization) problem in Equation (1). In particular, there may exist no single solution that can simultaneously optimize both of the above objectives (i.e., maximize speak rate and minimize source dependency) given the fact that they could be odds with each other. Therefore, the goal of the above multi-objective optimization problem is to identify the operation points along the Pareto frontier.

4. Source Selection

In the previous section, we formulate the *critical source selection* problem as a constraint optimization problem. One possible solution to the optimization problem is to perform a brute-force search. However, the time complexity of a brute-force search is $O(2^{|S|})$ ($|S|$ is the number of sources), which is not practical in many social sensing applciations. Therefore, we need to develop a more efficient solution. In the rest of this section, we first prove that the formulated critical source selection problem is NP-hard. We then present the details of our Critical Source Selection (CSS) scheme.

4.1. Complexity Analysis of the Formulated Problem

In this subsection, we prove that the formulated problem is a NP-hard problem. Based on the definitions in Section 3, we construct a graph $G = (S, C; E_S, E_{SC})$ based on the Source-Claim Matrix SC, Source-Dependency-Score Matrix SDS and Speak-Rate Vector SR as follows:

- A source S_i represents a vertex in S;
- A claim C_j represents a vertex in C;
- E_S is the set of edges between the vertices of S to represent the dependency between sources in SDS. In particular, if the element $SDS_{i,i'} > 0$, we have an edge between source S_i and $S_{i'}$.
- E_{SC} is the set of edges between vertices of S and C to represent report behaviors in SC. Specifically, if the element $SC_{i,j}$ in SC is 1, we have a edge between source S_i and claim C_j.
- We define wv_i as SR_i (i.e., speak rate of source S_i) to represent the vertex weight of vertex S_i and $we_{i,i'}$ as $SDS_{i,i'}$ (i.e., dependence score between source S_i and $S_{i'}$) to represent the edge weight between vertex S_i and $S_{i'}$.
- We further define two weight functions $w_{E_S} : E_S \mapsto R^+$ and $w_S : S \mapsto R^+$ to represent the dependency scores between sources and speak rates of sources respectively.

In particular, SR_i is the speak rate of source S_i defined in the previous section (i.e., $SR_i = \frac{\sum_{j=1}^{Y} SC_{i,j})}{Y}$). The objective is to find a subset S_c of S such that every vertex in C is connected to the vertex in S_c and satisfies the following objectives:

i) the sum of vertex weights in S_c (i.e., $\Sigma\ wv_i;\ S_i \in S_c$) is maximized;

ii) the sum of edge weights $we_{i,i'}$ in the subgraph induced by S_c (i.e., $\Sigma\ we_i;\ S_i, S_{i'} \in S_c$) is minimized;

We first consider a simplified version of the above problem by only considering the objective of minimizing the total dependency scores of selected sources. If we can prove that this simplified version is NP-hard, the original version is also NP-hard. We formally define the decision version of the simplified problem as follows:

Definition 4. *Given a graph $G = (S, C; E_S, E_{SC})$, a weight function $w_{E_S} : E_S \mapsto R^+$, a weight function $w_S : S \mapsto R^+$, and a positive number k, where S and C are two sets of vertices. E_S is a set of edges only among the vertices of S. E_{SC} is a set of edges between vertices S and C. The objective is to decide whether there is a subset S_c of S such that every vertex in C is connected to the vertex in S_c and the sum of edge weights in the subgraph induced by S_c is at most k.*

To prove that the simplified version is NP-hard, we need to demonstrate that the decision version is NP-complete. After that, we can conclude that the problem formulated in Equation (1) is NP-hard. The proof details are presented in the Appendix (Section 8).

4.2. The Critical Source Selection Scheme

The proof in above section shows that the formulated problem is NP-hard. We need to develop an efficient solution to select the critical set of sources. In this work, we propose the *Critical Source Selection (CSS)* scheme that consists of two major components: Candidate Source Selection and Annealing based Source Selection.

4.2.1. Candidate Source Selection

Based on the problem formulation in Equation (1), there are two objectives: (i) maximize the speak rates of the selected sources; (ii) minimize the dependency scores between the selected sources. We take a common approach in optimization and use linear combination to convert multi-objective programming to single-objective programming [45]. We can rewrite Equation (1) with

an objective function as:

$$f = max \sum_{i=1}^{X} SR_i \cdot \eta_i - \varphi \cdot \sum_{i=1}^{X} \sum_{i' \neq i} SDS_{i,i'} \cdot \eta_i \eta_{i'}$$

$$s.t. \;\; \eta_i \in \{0,1\}, \quad i = 1, ..., X$$

$$\bigcup C_{S_i} = C, \quad S_i = 1 \text{ and } S_i \in S \text{ where } \delta_i = 1 \qquad (2)$$

where φ is a parameter to balance our two objectives.

We denote a graph $G_s = (S, E_S, W_e, W_v)$, where W_e, W_v represent the set of edge weights and the set of vertex weights respectively. Without loss of generality, we use v_i, $e_{i,i'}$, $we_{i,i'}$, wv_i to represent the vertex, edge, edge weight and vertex weight respectively ($i \in [1, ..., X]$, $i' \in [1, ..., X]$ and $i \neq i'$). In this work, v_i is source S_i. $e_{i,i'}$ represents the dependency relationships between source S_i and $S_{i'}$. We further define Ne_i as the vertices which are connected to vertex v_i and t as the iteration index.

We first construct set S_c and C^* to contain the selected sources and the set of claims reported from the selected sources, respectively. The key steps of the Source Candidate Selection scheme are summarized as:

- We initialize S_c and C^* as $\emptyset$.
- We do the following three sub-steps iteratively:

  i) We select the vertex in graph G_s with the largest vertex weight wv_i. Without loss of generality, we denote the selected node as v_i.

  ii) We conduct vertex weight updates on other vertices which connected to vertex v_i. Specifically, the weight $wv_{i'}$ on vertex $v_{i'}$ ($i' \in Ne_i$) is updated as $wv_{i'} = wv_{i'} - we_{i,i'} \cdot \varphi$. Here, we update the vertex weight of the connected vertices of v_i by balancing the two objectives in Equation (2).

  iii) We add vertex v_i to the set of selected sources S_c and remove it from graph G_s together with all the edges connected to vertex v_i.

Figure 2 shows a simple illustrative example of the Candidate Source Selection scheme. In the source selection process, we firstly select vertex v_5 with the largest vertex weight. After that, we update the vertex weights of the vertices connected to v_5 and remove v_5 as well as the corresponding edges from the current graph.

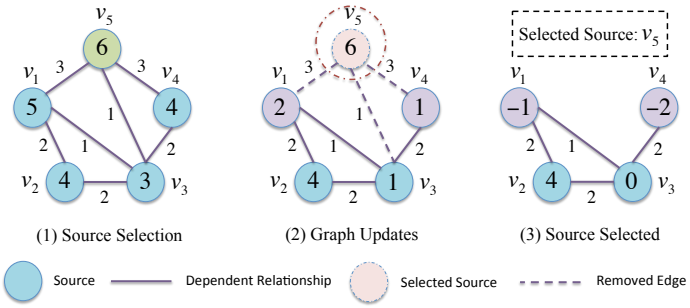


Figure 2: Simple Illustrative Example for Source Candidate Selection scheme

Algorithm 1: Source Candidate Selection Scheme

Input: A weighted and undirected graph $G_s = (S, E_S, W_e, W_v)$ and the full set of claim C.
Output: A set of selected sources S_c.
Initialize: $S_c^0 \leftarrow \emptyset$, $C_0^* \leftarrow \emptyset$, $G_s^0 \leftarrow G_s$, $E_S^0 \leftarrow E_S$, $t \leftarrow 0$
repeat
  Select the vertex in graph G_s with the largest vertex weight. (without loss of generality, we suppose that the selected node is v_i)
  for each $i' \in [1, ..., X]$ and $i' \neq i$ **do**
    $wv_i \leftarrow wv_i - we_{i,i'} \cdot \varphi$
    $S_c^{t+1} \leftarrow S_c^t \cup \{v_i\}$
    $G_s^{t+1} \leftarrow G_s^t - \{v_i\}$
    $E_S^{t+1} \leftarrow E_S^t - \{e_{i,i'}\}$
    $C_{t+1}^* = C_t^* \cup C_{S_i}$
  end for
  $t = t + 1$
until $C_t^* = C$

In summary, the input to the source candidate selection scheme is the generated graph G_s and the claims set C. The output of this scheme is the set

of critical sources S_c. The scheme is summarized in Algorithm 1. The time complexity of the first step (i.e., vertex selection) is of order $O(|S|)$ and the time complexity of the second step (i.e., vertex weight update) is also of order $O(|S|)$. We iteratively conduct the above two steps until $C^*_{t+1} = C^*_t \cup C_{S_i}$. Therefore, the time complexity of our source candidate selection algorithm is of order $O(|S| \cdot |S_c|)$, where $|S_c|$ is the size of selected critical source set. Since $|S_c|$ is normally much smaller than $|S|$, our scheme is scalalbe in large-scale social sensing applications.

4.2.2. Annealing based Source Selection

In this subsection, we further refine our critical source set using an annealing based scheme. In particular, we first partition all sources S into set S_c (candidate source set) and S_n (non-candidate source set). We then iteratively search for the approximate optimal solutions from S_n to replace sources in S_c to improve the value of the objective function. We introduce the following definitions to be used in our scheme:

Definition 5. ***Objective Difference*** Δf*: We define Objective Difference* Δf *to represent the difference between the values of objective functions from old and new solutions. Formally,* $\Delta f = f(S_c^{l+1}) - f(S_c^l)$*, where* $f(S_C^{t+1})$ *and* $f(S_c^t)$ *represent the value of objective function from the new* (S_c^{l+1}) *and old solution set* (S_c^l)*, respectively.* l *is the iteration index.*

Definition 6. ***Acceptance Probability Function*** AF*: We define Acceptance Probability Function* AF *to represent the probability of moving to the new solution from the current solution. Formally, we define* AF *as follows:*

$$AF = exp(\frac{f(S_c^{l+1}) - f(S_c^l)}{T}) \tag{3}$$

where T *represents the temperature parameter [46] which is a function of iteration index. We define* $T = |S_c| \times |S_n| - l$*.*

We observe that: (i) AF is larger than 1 if the new solution is better than the old one (i.e., $\Delta f > 0$). (ii) AF is smaller than 1 if the new solution is worse

than the old one (i.e., ($\Delta f < 0$). (iii) AF becomes smaller as the temperature T decreases when the new solution is worse than the old one.

In this paper, we propose an Annealing based Source Selection scheme to select the critical set of sources. Given the two sets: S_c and S_n, without loss of generality, we use u and v to represent the source selected from S_c and S_n, respectively. We denote the size of set S_c and S_n as U and V, respectively. We update S_c^l (l is the iteration index) iteratively based on the following key steps until the temperature parameter T decreases to 0.

- We calculate the objective function value $f(S_c^l)$ based on the current solution S_c^l.
- We select the source u in S_c with the smallest speak rate and select the source v in S_n with the largest speak rate. This selection process is based on the observation that sources with larger speak rate are more likely to achieve larger values in the objective function [43].
- We generate the new solution S_c^{l+1} by replacing u with v, and recalculate the value of objective function $f(S_c^{l+1})$ based on the new solution.
- We calculate the acceptance probability function AF and decide whether to move to the new solution from the current one or not. In particular, if $AF > 1 (i.e., \Delta f > 0)$, we move to the new solution S_c^{l+1} and consider it as the base for its next iteration. If $AF \leq 1$, we compare AF with a random number r which is between 0 and 1. If $AF > r$, we move to the new solution. Otherwise, we keep the current solutions S_c^l.

Figure 3 shows a simple illustrative example of the Annealing based Source Selection scheme. In summary, the input to the scheme is the candidate source set S_c and non-candidate source set S_n. The output of the scheme is the set of critical sources S^*. The time complexity of the annealing based source selection scheme is $O(|S_c| \times |S_n|)$ where $|S_c|$ and $|S_n|$ represents the size of S_c and S_n, respectively. We summarize the Annealing based Source Selection scheme in Algorithm 2.

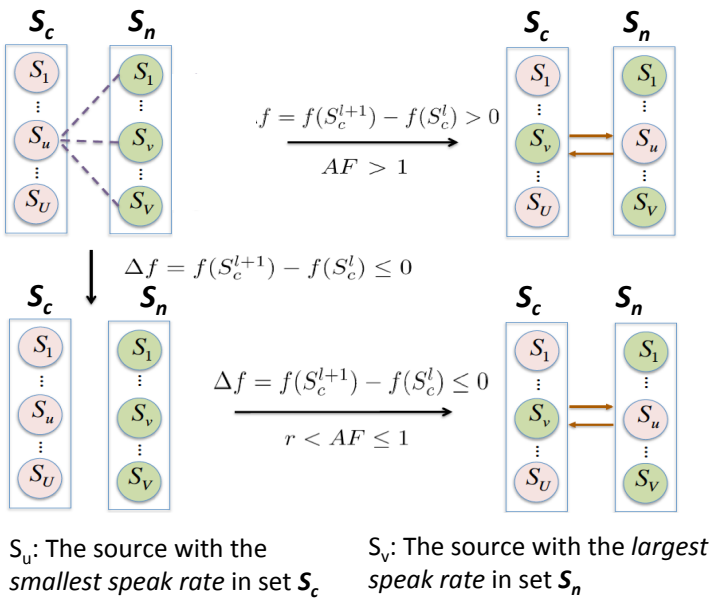


Figure 3: Simple Illustrative Example for Annealing based Source Selection Scheme

5. Evaluation

We carry out extensive experiments to demonstrate the efficacy and efficiency of the *CSS (Critical Source Selection)* scheme using the real-world datasets collected from a social sensing application. The evaluation results demonstrate that *CSS* can help to achieve better truth discovery results at a faster speed by judiciously selecting the critical set of sources.

5.1. Experimental Setups

5.1.1. Data Trace Statistics

In this paper, we evaluate our proposed scheme on two real-world data traces collected from Twitter in the aftermath of recent emergency and disaster events. Twitter is an open social sensing data collection platform where a large number of observations are contributed voluntarily by human sources to report on events in the physical environment. On Twitter, users have both explicitly (e.g., following relationship) and implicit (e.g., retweet behavior) dependencies, and

Algorithm 2: Annealing based Source Selection Scheme

Input: Candidate Source Set S_c **and Non-Candidate Source Set** S_n
Output: Critical Source Set S^*
B **is the set of replaced sources in** S_c **and** $S_c \setminus B = \{x \in S_c | x \notin B\}$
Initialize: $B \leftarrow \emptyset$, $T = |S_c| * |S_n|$, $l \leftarrow 0$
repeat
    Select the source in $S_c \setminus B$ with the smallest speak rate.
    Select the source in S_n with the largest speak rate.
    Calculate the value of objective function $f(S_c^l)$ based on Equation (2).
    Generate the new solutions S_c^{l+1} by replacing u with v and recalculate the value of objective function $f(S_c^{l+1})$.
    Calculate the acceptance probability function AF based on Equation (3).
    if $AF > 1$ **then**
        Move to the new solution S_c^{l+1}.
        Add u into set B.
    else if $AF > r$ **then**
        Move to the new solution S_c^{l+1}.
        Add u into set B.
    else
        Keep the current solutions S_c^l.
    end if
    $T = |S_c| * |S_n| - l$
    $l = l + 1$
until $T = 0$
$S^* \leftarrow S_c$

they tweet with different speak rates (e.g., some Twitter users are more chatty than other users). These relevant features of Twitter users provide us a good opportunity to evaluate the performance of the *CSS* scheme in real world social sensing scenarios. In the evaluation, we selected two data traces: (i) the Paris Attack event which happened on Nov, 2015; (ii) the Oregon Shooting which happened on Oct, 2015. Table 2 presents the statistics information of two data traces.

Table 2: The Statistics of Data Traces

Data Trace	Paris Attack	Oregon Shooting
Start Date	11/13/2015	10/1/2015
Duration	Eleven Days	Six Days
Physical Location	Paris, France	Umpqua, Oregon
Search Keywords	Paris, Attacks, ISIS	Oregon, Shooting, Umpqua
# of Tweets	873,760	210,028

5.1.2. Data Pre-Processing

To evaluate our proposed approach in real world scenarios, we go through the data pre-processing steps to generate the following inputs for the *CSS* scheme: (i) Source-Claim Matrix (i.e., SC Matrix); (ii) Speak-Rate Vector (i.e., SR Matrix); (iii) Source-Dependency-Score Matrix (i.e., SDS Matrix). We summarize the pre-processing steps as follows.

- *Source-Claim Matrix Generation*: We cluster similar tweets into the same cluster by leveraging an improved K-Means clustering algorithm and the Jaccard distance, which is a commonly used distance metric for social media data clustering [47]. Specifically, the Jaccard distance is defined to measure the distance (dissimilarity) between two tweets. It can be formally represented as $1 - \frac{\varphi_1 \cap \varphi_2}{\varphi_1 \cup \varphi_2}$ where φ_1 and φ_2 indicates the word set of the two compared tweets. In our experiments, we tokenize the tweets into individual words (removing special symbols and stopping words) and then compute the Jaccard distance between the pair of tweets. Then, we take each user as a source and each cluster as a claim in our model. We construct the SC matrix by associating each source with his/her reported claims. Specifically, the element $SC_{i,j} = 1$ in matrix SC if the tweet reported by source S_i belongs to cluster/claim C_j. Otherwise, we set $SC_{i,j} = 0$.

- *Speak-Rate Vector Generation*: We generated the SR Vector based on the constructed SC Matrix from the previous step. In particular, element SR_i in SR is the number of claims reported by source S_i normalized by the

total number of claims. Formally, $SR_i = \frac{\Sigma_{j=1}^{Y} SC_{i,j}}{Y}$.

- *Source-Dependency-Score Matrix Generation*: We construct the SDS Matrix based on source reporting behavior on Twitter. Particularly, we generate the source dependency graph as an undirected graph $G_{sds} = (V_{sds}, E_{sds}, W_{sds})$ where V_{sds} represents the set of sources, E_{sds} represents the set of their dependency links and W_{sds} represents the dependency degree between them. We use the following heuristic to construct the links in the graph G_{sds}: an undirected edge from source S_i to source $S_{i'}$ is added if there exists at least one claim reported by both source $S_{i'}$ and S_i. We then generate the Social-Dependency-Score Matrix SDS by setting the corresponding element $SDS_{i,i'}$ as the number of claims reported by both source S_i and $S_{i'}$. The reason we do not choose the follower-followee relationship to compute source dependency score are threefold: i) there exists a rate limit on Twitter to collect follower-followee data, so we cannot collect the complete dependency graph; ii) the follower-followee is more static and less appropriate to capture the dynamic dependency between sources on Twitter [7]; iii) it might raise some privacy concerns. The above heuristic approach is an approximation to estimate source dependency from social sensing data in real-world applications. We will further refine our techniques on source dependency graph estimation in the future.

5.1.3. Evaluation Metric

In our evaluation, we use the following metrics to evaluate the estimation performance of the *CSS* scheme: *Precision*, *Recall*, *F1-score* and *Accuracy*. Table 3 formally presented their definitions. In our evaluation, True Positives and True Negatives represent claims that are correctly classified by a particular scheme as true or undecided claims, respectively. The False Positives and False Negatives represent the true and undecided claims that are misclassified to each other.

Table 3: Metric Definitions

Metric	Definition
$Precison$	$\frac{TP}{TP+FP}$
$Recall$	$\frac{TP}{TP+FN}$
$F1-measure$	$\frac{2\times Precison\times Recall}{Precison+Recall}$
$Accuracy$	$\frac{TP+TN}{TP+TN+FP+FN}$

5.2. Evaluation of Our Methods

The state-of-the-art source selection baselines we used in the evaluation include:

- *DS* selects a set of diversified sources by only considering the dependencies among sources using a set of heuristic based approaches in social sensing applications [10].
- *FS* selects a set of sources by considering source freshness based on the source reporting behaviors [11].
- *LM* selects a set of sources by considering source speak rate for data integration.[12].
- *PS* selects the subset of data sources to predict the state of all other sources by considering source dependency in order to minimize the prediction errors on disaster response [13].

To evaluate all source selection schemes, we use the selection results from different algorithms as input to the state-of-the-art truth discovery techniques that include:

- *TASD* explores the topic relevance feature of claims and the arbitrary source dependencies among sources to ascertain the correctness of claims [6].

- *TD-C* solves the problem of truth discovery for correlated claims by modeling claims' correlations as regularization terms [48].

- *CAT* solves the truth discovery problem by considering the interdependence between claims and formulate it as a Bayesian network [3].

- *AS* solves the truth discovery problem by explicitly modeling the dependencies among sources on social networks using an estimation theoretic approach [7].

Even though the computational complexity of many truth discovery solutions is linear with respect to the number of sources, there also exist a non-trivial number of truth discovery schemes whose computational complexity is non-linear with respect to the number of sources [49]. Examples of these schemes include TASD [6] and AS [7], two of the truth discovery solutions we introduced above. Also, the number of sources can be quite large in real world social sensing applications (e.g., from tens of thousands to millions) [17]. Therefore, it makes sense to develop an effective source selection scheme to identify the critical sources and improve the computational efficiency of the truth discovery solutions.

In our evaluation, we combined each source selection scheme with different truth discovery techniques. We manually graded the output of these combinations to determine the correctness of the claims. Considering the manpower limitations, we took the union of the top 50 claims returned by different schemes as our evaluation set in order to avoid the bias towards any particular scheme. The following rubric was used to collect the ground truth information of the evaluation set:

- *True Claims*: Claims that are statements of an event, which is generally observable by multiple independent sources and can be corroborated by credible sources external to Twitter (e.g., mainstream news media).

- *Undecided Claims*: Claims that do not meet the criteria of true claims.

We note that undecided claims can potentially consist of two types of claims: (i) true claims that cannot be independently verified by external sources; (ii) false claims. Thus, our evaluation actually provides pessimistic performance bounds on estimations by treating undecided claims as false.

Table 4: Critical Source Selection Evaluation on Paris Attack Data Trace

Alg	Truth Discovery	Accuracy	Precision	Recall	F1-score
CSS-W/ Annealing	TASD	**0.706**	**0.802**	**0.767**	**0.784**
	TD-C	**0.640**	**0.702**	**0.840**	**0.765**
	CAT	**0.633**	**0.704**	**0.814**	**0.755**
	AS	**0.696**	**0.800**	**0.756**	**0.777**
CSS-W/O Annealing	TASD	0.683	0.802	0.724	0.761
	TD-C	0.632	0.704	0.814	0.755
	CAT	0.600	0.714	0.709	0.711
	AS	0.670	0.798	0.705	0.749
DS	TASD	0.494	0.654	0.578	0.614
	TD-C	0.541	0.702	0.593	0.643
	CAT	0.531	0.680	0.618	0.647
	AS	0.493	0.654	0.578	0.614
FS	TASD	0.556	0.784	0.502	0.612
	TD-C	0.572	0.695	0.665	0.680
	CAT	0.541	0.685	0.632	0.657
	AS	0.618	0.789	0.614	0.691
LM	TASD	0.562	0.786	0.510	0.618
	TD-C	0.577	0.695	0.698	0.696
	CAT	0.531	0.682	0.625	0.652
	AS	0.567	0.785	0.520	0.625
PS	TASD	0.602	0.797	0.574	0.668
	TD-C	0.564	0.695	0.665	0.680
	CAT	0.542	0.685	0.632	0.657
	AS	0.615	0.788	0.611	0.688
No Source Selection	TASD	0.592	0.787	0.563	0.656
	TD-C	0.569	0.696	0.676	0.686
	CAT	0.522	0.677	0.603	0.638
	AS	0.572	0.788	0.527	0.632

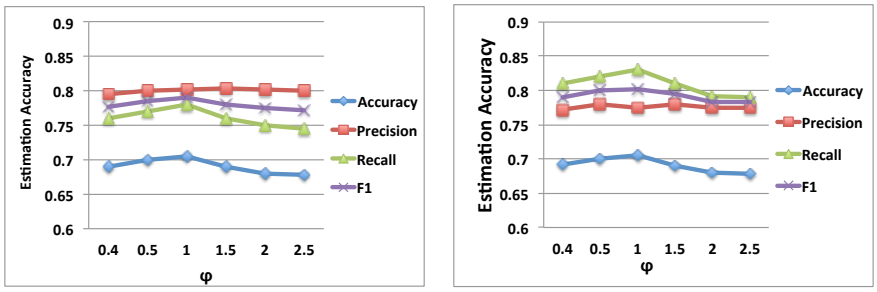
The evaluation results of Paris Attack data trace are shown in Table 4. We can observe that *CSS* scheme outperforms the compared baselines with different truth discovery techniques in all evaluation metrics. The largest performance

Table 5: Critical Source Selection Evaluation on Oregon College Shooting Data Trace

Alg	Truth Discovery	Accuracy	Precision	Recall	F1-score
CSS-W/ Annealing	TASD	**0.705**	**0.762**	**0.834**	**0.796**
	TD-C	**0.668**	**0.734**	**0.824**	**0.776**
	CAT	**0.685**	**0.757**	**0.812**	**0.783**
	AS	**0.691**	**0.764**	**0.808**	**0.785**
CSS-W/O Annealing	TASD	0.637	0.756	0.747	0.751
	TD-C	0.614	0.723	0.726	0.725
	CAT	0.637	0.728	0.743	0.735
	AS	0.654	0.757	0.746	0.752
DS	TASD	0.522	0.681	0.509	0.583
	TD-C	0.572	0.681	0.649	0.665
	CAT	0.562	0.673	0.644	0.658
	AS	0.519	0.683	0.495	0.574
FS	TASD	0.568	0.683	0.635	0.658
	TD-C	0.581	0.708	0.612	0.656
	CAT	0.596	0.741	0.589	0.656
	AS	0.556	0.676	0.616	0.645
LM	TASD	0.547	0.683	0.575	0.624
	TD-C	0.548	0.703	0.533	0.606
	CAT	0.562	0.720	0.570	0.636
	AS	0.523	0.679	0.514	0.585
PS	TASD	0.559	0.678	0.621	0.649
	TD-C	0.544	0.699	0.532	0.605
	CAT	0.602	0.753	0.584	0.657
	AS	0.544	0.673	0.588	0.628
No Source Selection	TASD	0.550	0.681	0.589	0.632
	TD-C	0.538	0.689	0.537	0.604
	CAT	0.537	0.670	0.573	0.618
	AS	0.544	0.715	0.505	0.592

gain achieved by *CSS* on F1-measure and accuracy over the best performed baseline (i.e., PS) are 11% and 10% respectively. The results of Oregon Shooting data trace are presented in Table 5. We can observe that *CSS* scheme continues to outperform all baselines with different truth discovery techniques. The performance improvements of *CSS* are achieved by explicitly considering both the source dependency and source speak rate in the sensor selection process, one of the main contributions of this paper.

Based on the evaluation results presented in Table 4 and 5, we observe that the "no source selection scheme" outperforms some source selection schemes in some metrics. The reason is: some source selection schemes aim to select a set of sources by considering only source freshness or source dependency. However, those selected sources may not be the critical sources for the truth discovery task in social sensing as we defined it in this paper. In particular, those schemes ignore the speak rate of sources. The speak rate has a strong positive correlation with both the accuracy and the granularity of the source reliability estimation, which also directly affects the claim correctness estimation [43]. Therefore, the non-critical sources selected by the above schemes might lead to truth estimation results that are worse than the scheme that uses all sources (i.e., no source selection).



(a) Paris Attack Data Trace (b) Oregon Shooting Data Trace

Figure 4: Evaluation Model Parameter φ Variation

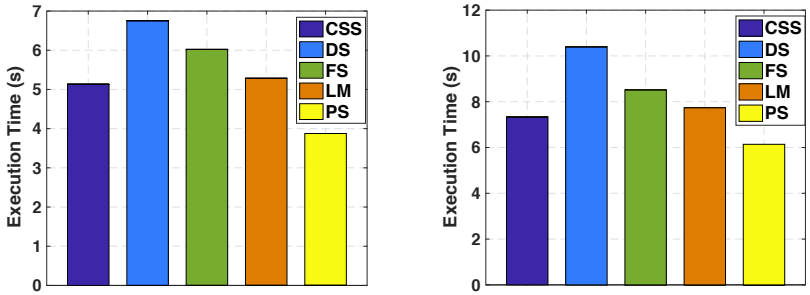
To investigate the effect of the parameter φ, which we used to update vertex weights in the source selection algorithm, we studied the performance of our proposed *CSS* scheme by varying the value of φ. Particularly, we vary the value of φ from 0.4 to 2.5. The evaluation results on Paris Attack and Oregon Shooting data traces are presented in Figure 4. We observe the performance of the CSS scheme degrades when φ is too small or too large. This again verifies the importance of considering both speak rate and source dependency in the source selection process for the truth discovery task. The optimal performance is achieved when $\varphi = 1$, which is used in our experiments. The CSS scheme

Table 6: Execution Time of Truth Discovery Schemes With/Without Source Selection

Alg	Truth Discovery Schemes	Execution Time (Seconds)
CSS	TASD	**181.2**
	TD-C	**32.73**
	CAT	**28.54**
	AS	**169.96**
No Source Selection	TASD	437.38
	TD-C	74.51
	CAT	67.86
	AS	400.34

becomes the LM scheme when φ is zero and the DS scheme when φ is infinity.

Finally, we evaluate the execution time of all compared algorithms on two real-world datasets. We run all algorithms on a regular lab computer (4 cores and 2 GHZ for each core, 8GB memory). Figure 5 presents the execution time of all algorithms on two real-world data traces. To highlight the differences, we only show the execution times of the source selection processes. We observe that the *CSS* scheme is among the fastest in all compared schemes, which demonstrates the efficiency of using *CSS* to identify the critical set of sources.



(a) Oregon Shooting Data Trace (b) Paris Attack Data Trace

Figure 5: Execution Time Comparison

We further perform experiments to evaluate the execution time of truth discovery schemes with and without the CSS scheme. The results on the Oregon Shooting trace are reported in Table 6. We observe that the CSS scheme greatly

reduces the execution time of the truth discovery algorithms by selecting the right sources. In particular, the reduction in the number of sources achieved by CSS is 2.37 and 2.42 on the Paris Shooting and Oregon Shooting datasets respectively, which explains the performance gain achieved by CSS. The results on the Paris Shooting trace are similar and we do not repeat them here.

6. Discussion

Notwithstanding the interesting problem and promising results reported in this paper, there exist a few directions for future work.

First, we mainly focus on the source dependency and speak rate features in the proposed CSS scheme. However, other features might also affect the source selection process. For example, if there exists some prior knowledge on the source reliability, that prior knowledge can be incorporated into the CSS scheme to guide the source selection scheme to choose more reliable sources. Moreover, the uncertainty on the truth discovery results [50] can also be used as a useful feature in the CSS scheme to identify sources with high confidence in their source reliability estimation. A potential challenge is to carefully incorporate these new features into the objective function of the source selection problem and to develop an efficient and optimized solution.

Second, we define a simple and effective source dependency score that depends on the number of commonly reported claims to measure the dependencies between sources in our model. However, such dependency score may not be perfectly accurate when two independent sources happen to make many common claims (e.g., two Twitter users who happen to appear at the same location around the same time during an event). To address this problem, we will develop more comprehensive metrics to accurately measure the complex dependencies between sources in our future work. For example, we plan to explore the techniques from information diffusion and epidemics [51, 52] to obtain a better understanding on how the information is propagated between sources and integrate such understanding into new metrics that can capture the complex

dependencies between sources in social sensing applications.

Third, the ideal case of source selection in truth discovery is to select the conditionally independent sources given the correctness of a claim. However, it is extremely difficult (if possible) to compute the actual conditional dependency scores in real-world social sensing applications because the correctness of claims is often unknown a priori. Therefore, we choose a statistical dependency measure in this work to identify independent sources. This choice is backed up by a few recent publications that also showed that selecting statistically independent sources could effectively improve the truth discovery accuracy in social sensing [43, 7]. We also acknowledge that there is a the truth discovery performance tradeoff between the corroboration between different sources on a claim and the statistical dependency among those sources. One goal of this paper is to explore this tradeoff and to identify a sweet point to improve the truth discovery performance. In future work, we will further explore more source dependency metrics to explore this tradeoff.

Fourth, in our critical source selection scheme, we consider both speak rate and source dependency in the source selection process. The Twitter accounts associated to various news outlets (e.g., CNN and BBC) might be the chattier than normal Twitter users. However, these Twitter accounts are also more likely to be followed and retweeted by other users on Twitter (hence a stronger source dependency than normal Twitter users). Our CSS scheme explores the tradeoff between speak rate and source dependency to identify critical sources (i.e., maximize speak rate and minimize source dependency), which will not rank the Twitter accounts associated to various news outlets higher than other sources. For example, the percentage of sources as news media outlets in the Paris Shooting dataset is 6.7% and 7.5% before and after the source selection process. The numbers for the Oregon Shooting dataset are 4.5% and 5.2% before and after the source selection process. We observed that: i) the percentage of sources that are news media sources is relatively small in both datasets; ii) our source selection algorithm does not have a significant bias of choosing the sources that are news media outlets over the common Twitter users. This is

again due to the fact we consider both the speak rate and source dependency in the CSS scheme. We did not remove the sources tied to news media outlets from evaluation because we want to keep the integrity of the datasets collected from the real world events.

Fifth, a chatty source might be picked up by our scheme if this source has little dependency with other sources. However, the source selection is only a filter for the truth discovery process, which is able to identify such chatty but untruthful source by estimating the source reliability [27]. The authors also plan to further enhance the CSS scheme by incorporating the source reliability (e.g., from prior estimation or external knowledge) into the source selection process. The challenge is how to solve the optimal source selection problem that considers source speak rate, dependency and source reliability under a unified framework.

Sixth, the source selection scheme proposed in this paper can be considered as a pre-screening phase of the truth discovery task when sources are not completely independent. An alternative approach to address non-independent sources is to explicitly model the source dependencies in the truth discovery models. The authors have done some preliminary work along this direction [7]. The key technical challenge is to keep the rigorousness of the analytical model of truth discovery tasks while accurately modeling the arbitrary dependencies between sources. It would also be interesting to explicitly compare the performance of the source selection (i.e., prescreening) based approach with the approach that explicitly models the source dependency in truth discovery tasks. It is also possible to develop a hybrid solution that combines the above two approaches. The authors plan to pursue this direction in their future work.

Finally, we note that the source selection scheme developed in this work is not only applicable to applications based on Twitter. It can be also applied to many other truth discovery applications where data is contributed by possibly dependent sources. Selecting the critical sources in these applications is an important task because it can not only effectively improve the accuracy of the truth discovery tasks but also improve the computational efficiency. In our future work, we plan to apply our CSS scheme in other truth discovery appli-

cations beyond Twitter (e.g., Location-based Social Networks (LBSNs), mobile crowdsensing applications) to further evaluate the performance and robustness of our scheme in different application scenarios.

7. Conclusion

In this paper, we develop a new critical source selection in social sensing to effectively reduce the complexity of a truth discovery problem and improve the accuracy of estimation results at the same time. In particular, our proposed scheme (CSS) explicitly explores source dependency and speak rate in the solution of critical source selection. We perform extensive experiments to compare the performance of our CSS scheme with state-of-the-art baselines using real-world social sensing datasets. The evaluation results demonstrate the effectiveness and efficiency achieved by our scheme.

Acknowledgments

This research is supported in part by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795 and Army Research Office under Grant W911NF-17-1-0409. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] A. Guéziec, Crowd sourced traffic reporting, uS Patent 8,718,910 (May 6 2014).

[2] P. Marks, Crowds point out potholes on a map to speed up street repairs (2013).

[3] S. Wang, L. Su, S. Li, S. Hu, T. Amin, H. Wang, S. Yao, L. Kaplan, T. Abdelzaher, Scalable social sensing of interdependent phenomena, in: International Conference on Information Processing in Sensor Networks (IPSN), ACM, 2015, pp. 202–213.

[4] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, Z. Su, Understanding retweeting behaviors in social networks, in: International on Conference on Information and Knowledge Management (CIKM), ACM, 2010, pp. 1633–1636.

[5] G.-J. Qi, C. C. Aggarwal, J. Han, T. Huang, Mining collective intelligence in diverse groups, in: International Conference on World Wide Web (WWW), ACM, 2013, pp. 1041–1052.

[6] C. Huang, D. Wang, Topic-aware social sensing with arbitrary source dependency graphs, in: International Conference on Information Processing in Sensor Networks (IPSN), ACM/IEEE, 2016, pp. 1–12.

[7] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, et al., Using humans as sensors: an estimation-theoretic perspective, in: International Conference on Information Processing in Sensor Networks (IPSN), ACM/IEEE, 2014, pp. 35–46.

[8] R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, D. Srivastava, Fusing data with correlations, in: International Conference on Management of Data (SIGMOD), ACM, 2014, pp. 433–444.

[9] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, J. Han, A confidence-aware approach for truth discovery on long-tail data, VLDB Endowment (2014) 425–436.

[10] M. Y. S. Uddin, M. T. Al Amin, H. Le, T. Abdelzaher, B. Szymanski, T. Nguyen, On diversifying source selection in social sensing, in: International Conference on Networked Sensing Systems (INSS), IEEE, 2012, pp. 1–8.

[11] T. Rekatsinas, X. L. Dong, D. Srivastava, Characterizing and selecting fresh data sources, in: International Conference on Special Interest Group on Management of Data (SIGMOD), ACM, 2014, pp. 919–930.

[12] X. L. Dong, B. Saha, D. Srivastava, Less is more: Selecting sources wisely for integration, in: Proceedings of the VLDB Endowment, Vol. 6, VLDB Endowment, 2012, pp. 37–48.

[13] M. Hosseini, N. Nagibolhosseini, A. Barnoy, P. Terlecky, H. Liu, S. Hu, S. Wang, T. Amin, L. Su, D. Wang, et al., Joint source selection and data extrapolation in social sensing for disaster response, arXiv preprint arXiv:1512.00500.

[14] D. Wang, K. Qiu, L.-c. Wang, Design of dba algorithm in epon upstream channel in support of sla, JOURNAL-CHINA INSTITUTE OF COMMUNICATIONS 26 (6) (2005) 87.

[15] D. Wang, H. Ahmadi, T. Abdelzaher, H. Chenji, R. Stoleru, C. C. Aggarwal, Optimizing quality-of-information in cost-sensitive sensor data fusion, in: Distributed Computing in Sensor Systems and Workshops (DCOSS), 2011 International Conference on, IEEE, 2011, pp. 1–8.

[16] C. Huang, D. Wang, Critical source selection in social sensing applications, International Conference on Distributed Computing in Sensor Systems (DCOSS).

[17] C. C. Aggarwal, T. Abdelzaher, Social sensing, in: Managing and Mining Sensor Data, Springer, 2013, pp. 237–297.

[18] Y. Zheng, T. Liu, Y. Wang, Y. Zhu, Y. Liu, E. Chang, Diagnosing new york city's noises with ubiquitous data, in: International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp), ACM, 2014, pp. 715–725.

[19] D. Wang, T. Abdelzaher, L. Kaplan, Surrogate mobile sensing, IEEE Communications Magazine 52 (8) (2014) 36–41.

[20] J.-w. WANG, D. WANG, K. TIMO, Y.-p. ZHAO, A novel anti-collision protocol in multiple readers rfid sensor networks [j], Chinese Journal of Sensors and Actuators 8 (026).

[21] A. Mukherjee, V. Venkataraman, B. Liu, N. S. Glance, What yelp fake review filter might be doing?, in: International Conference on Web and Social Media (ICWSM), AAAI, 2013.

[22] P. Zhou, Y. Zheng, M. Li, How long to wait?: predicting bus arrival time with mobile phone based participatory sensing, in: International conference on Mobile systems, applications, and services (Mobisys), ACM, 2012, pp. 379–392.

[23] S. S. Kanhere, Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces, in: International Conference on Distributed Computing and Internet Technology (ICDCIT), Springer, 2013, pp. 19–26.

[24] S. Reddy, D. Estrin, M. Srivastava, Recruitment framework for participatory sensing data collections, in: Proceedings of the 8th International Conference on Pervasive Computing, Springer Berlin Heidelberg, 2010, pp. 138–155.

[25] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in: International Conference on World Wide Web (WWW)), ACM/IEEE, 2010, pp. 851–860.

[26] H. To, G. Ghinita, L. Fan, C. Shahabi, Differentially private location protection for worker datasets in spatial crowdsourcing, Transactions on Mobile Computing (TMC) 16 (4) (2017) 934–949.

[27] C. Huang, D. Wang, N. Chawla, Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems, Transactions on Big Data.

[28] D. Wang, M. T. Al Amin, T. Abdelzaher, D. Roth, C. R. Voss, L. M. Kaplan, S. Tratz, J. Laoudi, D. Briesch, Provenance-assisted classification

in social networks, IEEE Journal of Selected Topics in Signal Processing 8 (4) (2014) 624–637.

[29] D. Y. Zhang, D. Wang, H. Zheng, X. Mu, Q. Li, Y. Zhang, Large-scale point-of-interest category prediction using natural language processing models, in: Big Data (Big Data), 2017 IEEE International Conference on, IEEE, 2017.

[30] J. Marshall, D. Wang, Mood-sensitive truth discovery for reliable recommendation systems in social sensing, in: Proceedings of the 10th ACM Conference on Recommender Systems, ACM, 2016, pp. 167–174.

[31] M. T. Al Amin, T. Abdelzaher, D. Wang, B. Szymanski, Crowd-sensing with polarized sources, in: Distributed Computing in Sensor Systems (DCOSS), 2014 IEEE International Conference on, IEEE, 2014, pp. 67–74.

[32] J. Teng, B. Zhang, X. Li, X. Bai, D. Xuan, E-shadow: Lubricating social interaction using mobile phones, Transactions on Computers (TOC) 63 (6) (2014) 1422–1433.

[33] R. W. Ouyang, L. Kaplan, P. Martin, A. Toniolo, M. Srivastava, T. J. Norman, Debiasing crowdsourced quantitative characteristics in local businesses and services, in: International Conference on Information Processing in Sensor Networks (IPSN), ACM/IEEE, 2015, pp. 190–201.

[34] Z. Zhao, J. Cheng, W. Ng, Truth discovery in data streams: A single-pass probabilistic approach, in: International Conference on Conference on Information and Knowledge Management (CIKM), ACM, 2014, pp. 1589–1598.

[35] D. Y. Zhang, R. Han, D. Wang, C. Huang, On robust truth discovery in sparse social media sensing, in: Big Data (Big Data), 2016 IEEE International Conference on, IEEE, 2016, pp. 1076–1081.

[36] D. Wang, T. Abdelzaher, L. Kaplan, C. C. Aggarwal, Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications, in: The 33rd International Conference on Distributed Computing Systems (ICDCS'13), 2013.

[37] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, X. Mu, G. Madey, C. Huang, Towards scalable and dynamic social sensing using a distributed computing framework, in: Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on, IEEE, 2017, pp. 966–976.

[38] D. Y. Zhang, D. Wang, Y. Zhang, Constraint-aware dynamic truth discovery in big data social media sensing, in: 2017 IEEE International Conference on Big Data (IEEE BigData 2017), IEEE, 2017.

[39] H. Amintoosi, S. S. Kanhere, M. Allahbakhsh, Trust-based privacy-aware participant selection in social participatory sensing, Journal of Information Security and Applications 20 (2015) 11–25.

[40] Y. Zhuang, K. Wang, W. Wang, H. Hu, A hybrid sensing approach to mobile robot localization in complex indoor environments, International Journal of Robotics and Automation 27 (2) (2012) 198.

[41] E. Niforatos, A. Vourvopoulos, M. Langheinrich, P. Campos, A. Doria, Atmos: a hybrid crowdsourcing approach to weather estimation, in: International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, ACM, 2014, pp. 135–138.

[42] N. Subari, J. Mohamad Saleh, A. Y. Md Shakaff, A. Zakaria, A hybrid sensing approach for pure and adulterated honey classification, Sensors 12 (10) (2012) 14022–14040.

[43] D. Wang, L. Kaplan, H. Le, T. Abdelzaher, On truth discovery in social sensing: A maximum likelihood estimation approach, in: International Conference on Information Processing in Sensor Networks (IPSN), 2012.

[44] R. W. Ouyang, M. Srivastava, A. Toniolo, T. J. Norman, Truth discovery in crowdsourced detection of spatial events, Transactions on Knowledge and Data Engineering (TKDE) 28 (4) (2016) 1047–1060.

[45] H.-J. Zimmermann, Fuzzy programming and linear programming with several objective functions, Fuzzy sets and systems 1 (1) (1978) 45–55.

[46] P. J. Van Laarhoven, E. H. Aarts, Simulated annealing, in: Simulated Annealing: Theory and Applications, Springer, 1987, pp. 7–15.

[47] K. D. Rosa, R. Shah, B. Lin, A. Gershman, R. Frederking, Topical clustering of tweets, in: International Conference on Special Interest Group on Information Retrieval (SIGIR), ACM, 2011.

[48] C. Meng, W. Jiang, Y. Li, J. Gao, L. Su, H. Ding, Y. Cheng, Truth discovery on crowd sensing of correlated entities, in: International Conference on Embedded Networked Sensor Systems (Sensys), ACM, 2015, pp. 169–182.

[49] D. Wang, T. Abdelzaher, L. Kaplan, Social Sensing: Building Reliable Systems on Unreliable Data, Morgan Kaufmann, 2015.

[50] D. Wang, L. Kaplan, T. Abdelzaher, C. C. Aggarwal, On credibility tradeoffs in assured social sensing, IEEE Journal On Selected Areas in Communication (JSAC).

[51] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information diffusion through blogspace, in: Proceedings of the 13th international conference on World Wide Web, ACM, 2004, pp. 491–501.

[52] J. Yang, S. Counts, Predicting the speed, scale, and range of information diffusion in twitter., ICWSM 10 (2010) 355–358.

8. Appendix

8.1. Proof of the Formulated Problem

In this section, we proof that the defined decision version of the simplified problem in Definition 4 is NP-complete as follows:

Theorem 1. *The decision problem in Definition 4 is NP-complete.*

Proof. Firstly, we prove that the decision problem is NP. Given any $S' \in S$, we can verify whether each vertex in C has an edge with the vertex in S' and whether $|S'| \leq k$ in polynomial time. Hence, the decision problem is NP. Next, we prove that the decision problem is NP-hard.

Given any instance I of the set cover problem has the following form:

- a set $E = \{e_1, e_2, \ldots, e_m\}$;
- a collection S of sets $s_1, s_2, \ldots, s_n$ such that $s_i \in E$ $(i = 1, 2, \ldots, n)$;
- a parameter η such that we are supposed to answer whether there is a subset $S^* \in S$ such that $|S^*| \leq \eta$ and all the elements in E are covered by the sets in S^*.

We reduce this instance I to an instance I' of the decision problem as follows:

- for each set $s_i \in S$, we construct a vertex $u_i \in S$ to construct the set S;
- for each element $e_j \in E$, we construct a vertex $v_j \in C$ to construct the set C;
- if $e_i \in s_i$, we generate an edge between v_i and u_j to construct the set E_{SC};
- for each pair of vertices in S, we generate an edge between them to construct the set E_S;
- let the function w_{E_S} be $w_E(e) = 1$ for $\forall e \in E_S$;
- let the function w_S be $w_S(u_i) = 1$ for $\forall u_i \in S$;
- let $k = \binom{\eta}{2}$

This construction can be done in polynomial time.

If there is a subset $S^* = \{s_{i_1}, s_{i_2}, \ldots, s_{i_{\eta'}}\}$ for the instance I such that all the elements in E are covered and $\eta' \leq \eta$, in the instance I', we can choose the S^* as $\{u_{i_1}, u_{i_2}, \ldots, u_{i_{\eta'}}\}$. In such case, each vertex in S has an edge with the vertex in S^*, and the sum of edge weights in the subgraph induced by S^* is exactly $\binom{\eta'}{2}$ ($\leq \binom{\eta}{2}$). Thus, there is a solution to the decision problem.

If there is a subset $S^* = \{u_{i_1}, u_{i_2}, \ldots, u_{i_{\eta'}}\}$ such that each vertex in C has an edge with the vertext in S^*, and the sum of edge weights in the induced subgraph of S^* is no more than $\binom{\eta}{2}$, we can choose the sets $s_{i_1}, s_{i_2}, \ldots, s_{i_{\eta'}}$ as S^* for the instance I. Since the induced subgraph is a complete graph and each edge has weight with 1, we have $\binom{\eta'}{2} \leq \binom{\eta}{2}$, i.e., $\eta' \leq \eta$. All elements in E are covered by S^* according to our construction of S^*. Hence, S^* is a solution to the set cover problem. Therefore, we proved that the decision problem is NP-complete. As stated above, we conclude that the problem formulated in Equation (1) is NP-hard. □